Using Multilevel Models for Small Area Estimation

Ray Chambers and Nikos Tzavidis

Southampton Statistical Sciences Research Institute

University of Southampton

NCRM Summer School, 5 July 2005

Introduction

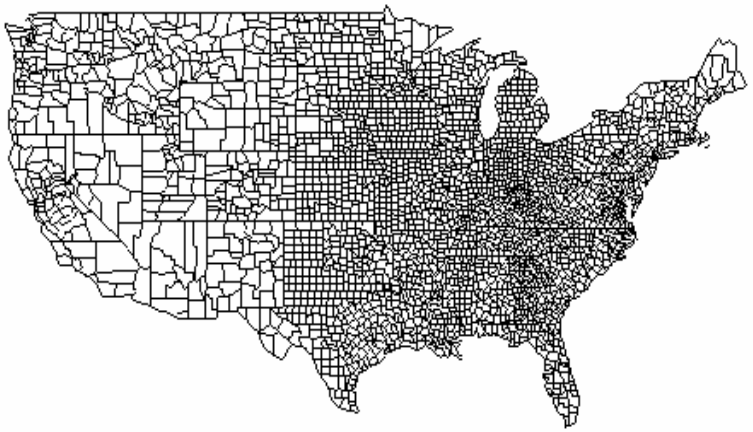
- Sample surveys are widely used to provide estimates of averages and other parameters for large populations and also for smaller sub-populations (domains)

- Domains often correspond to

  Geographic domains (areas): State, county, municipality, school district, health service area

  Socio-demographic groups: E.g. age by gender by race group in a large geographic area

Example: U.S. Geographic Domains (States and Counties)



Demand for Domain Statistics

The demand for domain statistics has increased due to their use in

- formulating policies
- allocation of government funds
- regional planning
- business decision making (e.g. many small businesses rely on information about local socio-economic conditions)

Applications of Small Area Estimation

Unemployment Rates: Estimates of unemployment levels and rates for U.K. Local Authority Districts have been recently produced by the Office for National Statistics

Poverty mapping: County level estimates of poor school-age children in the U.S. are used to allocate $7 billion for programs aimed at educationally disadvantaged children

Health: Local area estimates of disability, alcohol and drug use and rates of diseases are used to plan resource allocation for treatment needs

The Small Area Estimation Problem

- Estimates derived using only the area-specific data are known as **“direct estimates”**
- An area is regarded as **large (small)** if the sample drawn from the area is large enough/not large enough to yield **“direct estimates”** of adequate precision
- A large enough overall sample size to support **“direct estimation”** for all areas of interest rarely exists because most survey designs are optimised for accurate national, not sub-national, estimates
- Survey users typically require more detailed later analyses

A Solution to the Small Area Problem

- Modern small area estimation is based on **model-based methods**

- The idea is to use statistical models to link the variable of interest with supplementary contextual information, e.g. census and administrative data, for the small areas. The contextual information is assumed to explain part of the between area variability

- This implies that even if data are only available for a limited number of individuals in a small area, we have covariate information for all individuals in that area

- **Multilevel models** can be used for this purpose

Borrowing Strength

- The philosophy behind **multilevel models** is to include **random area effects** to account for between area variation beyond that explained by the covariate information

- A **random area effect** is a unique area-specific parameter that shows how different one area is from another

- Estimating the random effect for a particular area requires using data from all areas and not just the data of the particular area. This is known as “borrowing strength”

- The fact that we make use of all available data leads to an increase of the **effective sample size** for the particular area

Using Multilevel Models for Small Area Estimation

We are interested in estimating the mean, the total and percentiles of a variable Y in the small areas of interest

In general, a multilevel model for Y has the following form

Y = Fixed Part + Area Effect + Residual

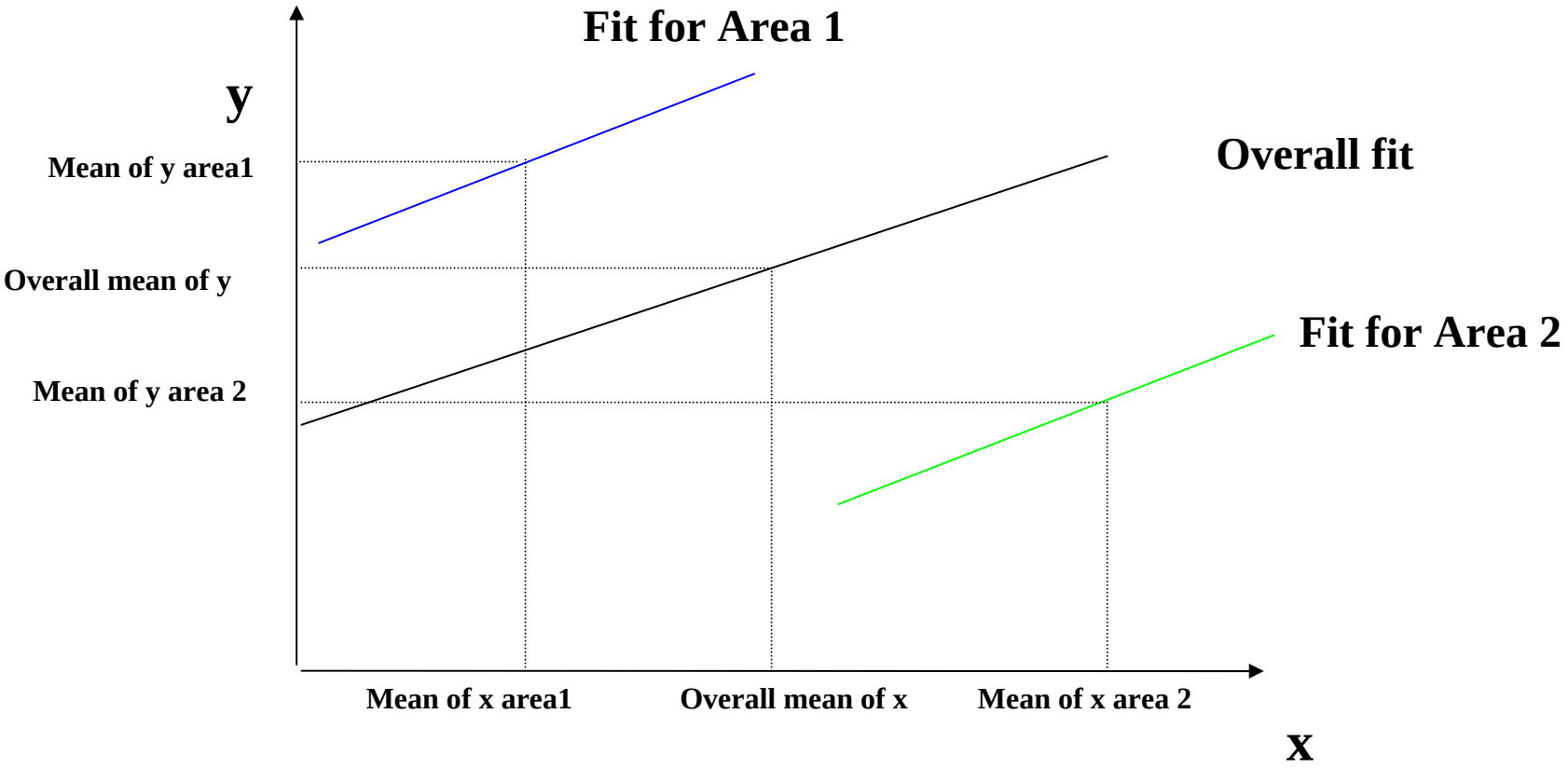
Fixed Part: Contribution of the covariate information to explaining between area variability

Area Effect: Area specific value that accounts for between area variability beyond that explained by the covariate information

Residual: Additional variability not explained by the model

Using Multilevel Models for Small Area Estimation (Contd)

Fitted lines for two areas estimated using a multilevel model



An Application: Labour Force Status Estimation for British LADs/UAs

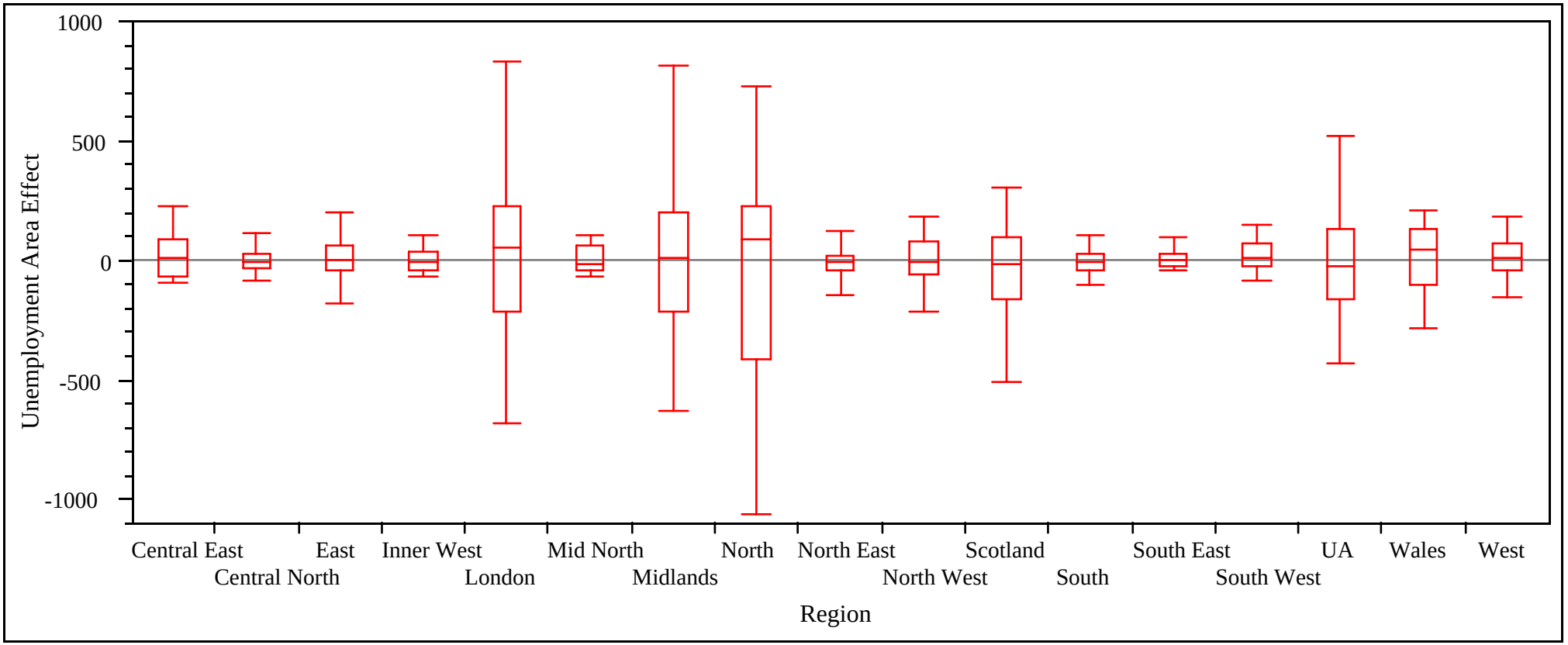
Logistic models for ILO unemployment and ILO inactive at age-sex group (6 categories) by LAD/UA level (402 areas)

Model

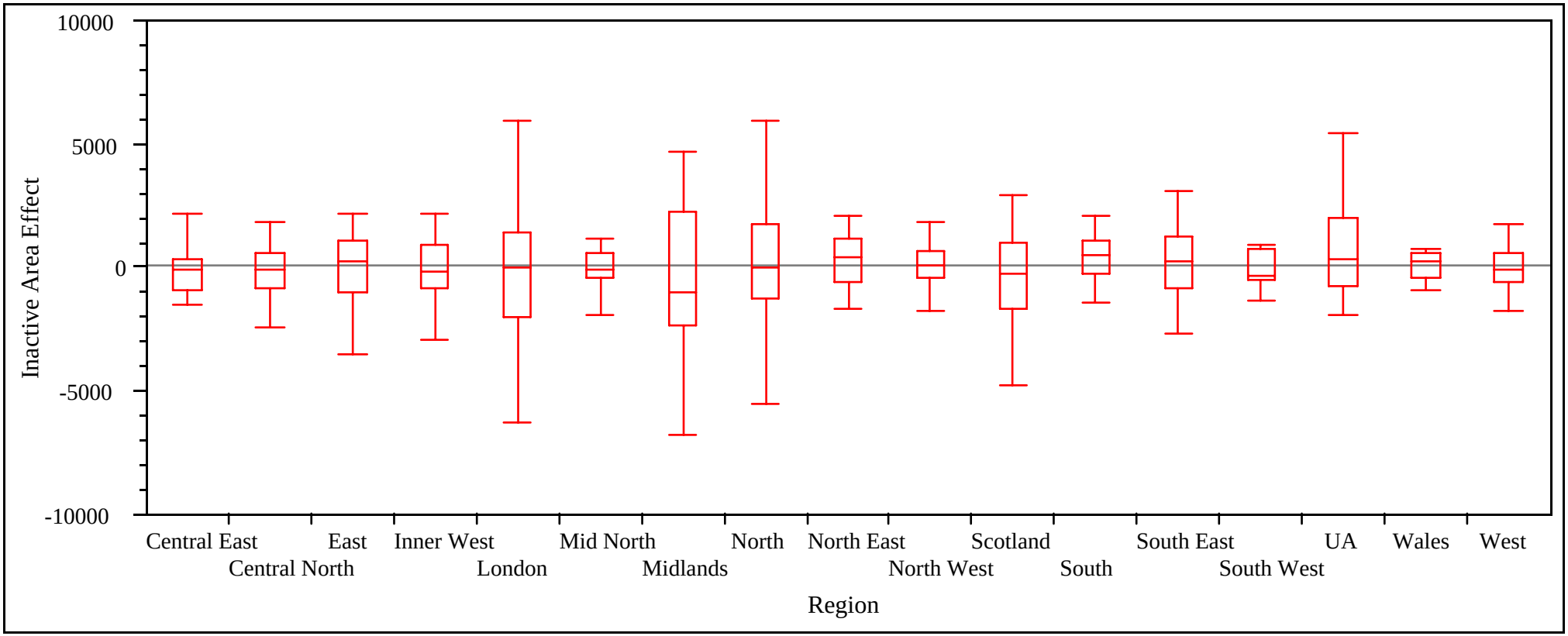
Fixed Part Age-sex by unemployment claimant count + region + socio-economic group + income support claimant counts (age, sex) + incapacity benefits claimant counts (age, sex)

Random Part LAD/UA-specific random effect

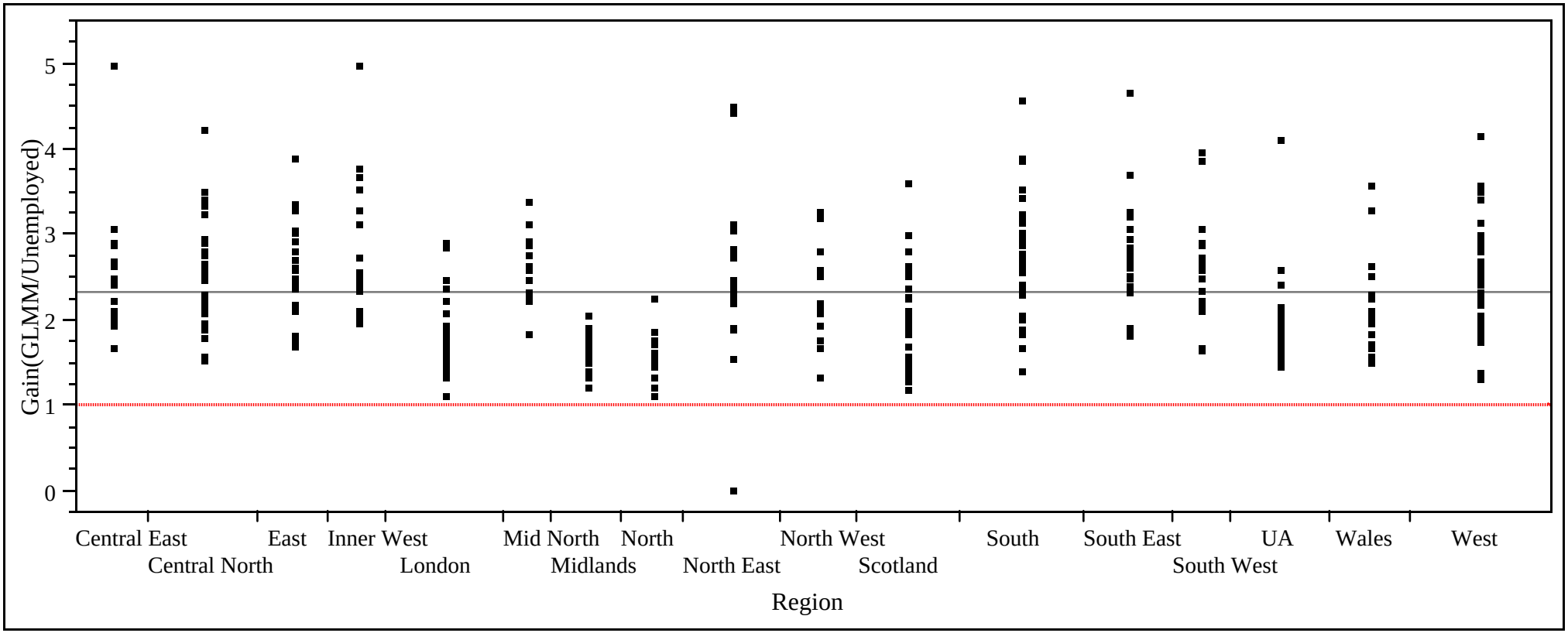
Distribution of Estimated LAD/UA Effects by Region for Unemployed



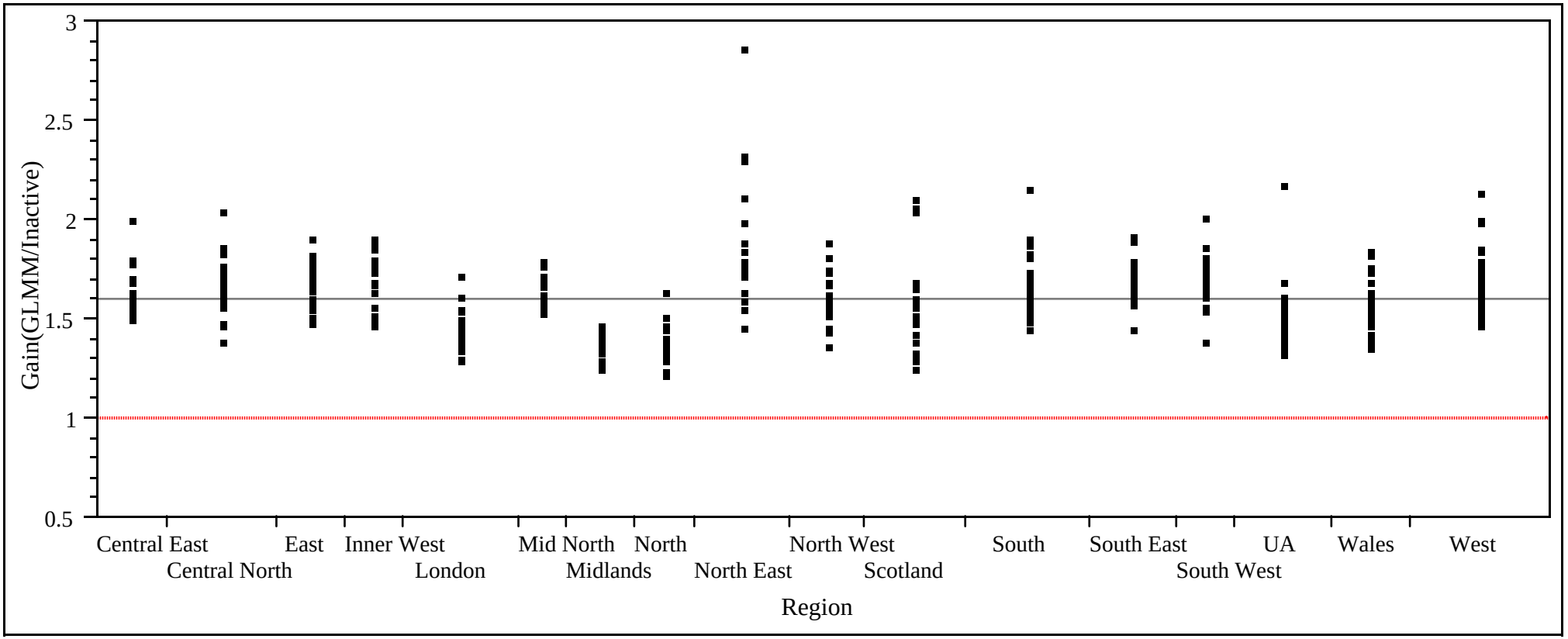
Distribution of Estimated LAD/UA Effects by Region for Inactive



Gains from Multilevel Logistic Model-Based Estimates for Unemployed



Gains from Multilevel Logistic Model-Based Estimates for Inactive



Summary

- Small area (domain) estimation is an important tool for decision making
- Reliable direct estimation of small area characteristics is difficult due to insufficient area-specific data
- A solution to this problem is to “borrow strength” from the different areas. One way to do this is via multilevel models
- This can lead to significant increases in effective sample size, and small area estimates of acceptable precision then become feasible